DSNDM: Deep Siamese Neural Discourse Model with Attention for Text Pairs Categorization and Ranking

Alexander Chernyavskiy and **Dmitry Ilvovsky**
National Research University Higher School of Economics
Moscow, Russia
alschernyavskiy@gmail.com; dilvovsky@hse.ru

Abstract

In this paper, the utility and advantages of the discourse analysis for text pairs categorization and ranking are investigated. We consider two tasks in which discourse structure seems useful and important: automatic verification of political statements, and ranking in question answering systems. We propose a neural network based approach to learn the match between pairs of discourse tree structures. To this end, the neural TreeLSTM model is modified to effectively encode discourse trees and DSNDM model based on it is suggested to analyze pairs of texts. In addition, the integration of the attention mechanism in the model is proposed. Moreover, different ranking approaches are investigated for the second task. In the paper, the comparison with state-of-the-art methods is given. Experiments illustrate that combination of neural networks and discourse structure in DSNDM is effective since it reaches top results in the assigned tasks. The evaluation also demonstrates that discourse analysis improves quality for the processing of longer texts.

1 Introduction

The growing popularity of social networks and the widespread use of social media contributed to the emergence of many NLP tasks associated with the processing of statements. It can be analyzed from an emotional point of view (sentiment analysis), opinion and argumentation mining, text summarization and so forth.

Despite the success of the transformer-based neural networks, such as BERT (Devlin et al., 2018) and its modifications, in various NLP tasks, they also have disadvantages since they frequently analyze only the plain text that can be quite long and complex. At the same time, discourse structure contains important knowledge for solving these tasks, and several researchers demonstrated its significance (Galitsky et al., 2015; Bhatia et al., 2015; Ji and Smith, 2017). However, the value of discourse have been already investigated only for some single text categorization tasks.

In this study, we demonstrate the utility and advantages of the matching of discourse tree structures of text pairs. Discourse analysis seems effective in textual entailment, text simplification and paraphrase detection tasks. However, it is necessary to analyze texts on the sentence level in most cases. We consider typical NLP tasks in which input texts are quite long and paragraphs are given initially.

One of such tasks is automatic verification of factual texts. Politicians may utilize unreliable statements for their own purposes. Due to the fact that there are plenty of such statements, it should be automatically evaluated for reliability and the possibility of manipulation of public opinion. In most cases, it is possible to extract some confirmation or refutation for a given factual text. In this way, we investigate the utility of discourse analysis in the classification of pairs of texts: statements and their justifications. Discourse structure may contain crucial knowledge even for the classification of the statements alone but can be even more effective in the case of analyzing additionally the confirmations and refutations.

Apart from that, one of the most appropriate tasks is the ranking in question answering systems. It was shown that discourse structure of questions and correct answers should correlate (Galitsky et al., 2015). Companies are interested in QA systems development in order to maximize the ease of interaction with customers. All questions can be divided into two groups: factoid and non-factoid. It is important to answer factoid questions to provide some specific information and non-factoid ones to maintain a dialogue. It is worth to emphasize that the second task is more challenging because there is no single correct answer for each question. We

Proceedings of the First Workshop on Computational Approaches to Discourse, pages 76–85
Online, November 20, 2020. ©2020 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17

consider the non-factoid questions asked on Internet forums since the discourse analysis seems to be more helpful in this case.

The main technical idea of this paper is to combine discourse analysis and recursive neural network TreeLSTM (Tai et al., 2015), which previously obtained the state-of-the-art results in some single text classification tasks.

Our contributions can be formulated as the following:

- We propose a neural network approach to learn the match between pairs of the discourse tree structures. To this end, we modify the basic TreeLSTM model to effectively encode the discourse structure and propose DSNDM (Deep Siamese Neural Discourse Model) to analyze pairs of texts.
- We suggest the way of integration of the attention mechanism in the DSNDM model.
- We investigate the value of the proposed approach considering two tasks and experimentally confirm the utility and importance of discourse analysis for the text pairs processing.

Our paper is organized as follows. Firstly, we summarize related work and introduce some base concepts. We continue with the description of the base model and its modifications. Then, we discuss the obtained results, error analysis and propose directions for further research.

2 Related Work

There are several approaches to solve the fact-checking problem. The best models presented in the FEVER competition (Thorne et al., 2018) allocate a stage of extracting supporting or refuting information and a classification stage. Justifications have already been extracted in our case. Therefore, there is no need to use the first stage. The BERT model (Devlin et al., 2018) is frequently used as the main model of the approach (Nie et al., 2019; Alonso-Reina et al., 2019). It is worth to emphasize that BERT cannot process long texts (the sequence is limited to 512 tokens). Therefore, it is necessary to extract the key information from the given justification paragraph. Besides, BERT can not efficiently store and process discourse features.

Another approach uses knowledge graphs. Clancy et al. (2019) proposed the use of relations between the entities of the graph in order to confirm some "distill" information extracted from the statement. Ciampaglia et al. (2015) suggested Knowledge Linker, the main idea of which is that if the path between entities in the knowledge graph is short, then the factual text containing them is reliable. It should be mentioned that this approach is generally applicable only to factoid statements since the entities must exist within knowledge graphs.

Finally, we distinguish the third approach which considers structural information extracted from texts (Wu et al., 2017; Galitsky and Ilvovsky, 2016). Galitsky et al. (2015) proposed to match discourse trees and solved the categorization task using Tree Kernel-based SVM. However, this approach does not utilize any modern neural networks. At the same time, recursive neural networks are gaining popularity (Ji and Smith, 2017; Bhatia et al., 2015; Tai et al., 2015). The main goal of them is to encode tree-like structures, such as syntax and discourse trees. These models achieved superior results in the single text categorization tasks, but researchers did not investigate the value of discourse analysis for processing pairs of texts. However, this approach is promising for the assigned task since frequently not only unreliable texts have a similar discourse structure, but the discourse structure of texts refuting them is also similar.

The main baselines for the question-answering problem are models that utilize keywords for ranking: using TF-iDF, BM25 and its modifications (Okapi BM25, BM25F). Frequently, their results are bad enough and need to be re-ranked using more complex methods. Neural network models, such as BERT, allow obtaining state-of-the-art results (Hashemi et al., 2020). It should be mentioned here that different training techniques of ranking are often not investigated.

In addition, some fact-checking approaches can be applied in question answering systems. For instance, Cui et al. (2017) and Liu et al. (2019) considered the possibility of using knowledge graphs. Galitsky (2019) investigated the value of discourse analysis in QA systems, but did not utilize any neural network approaches.

3 Methods

3.1 Discourse Tree Structure

Any text can be represented as a tree using the Rhetorical Structure Theory (RST) proposed by Mann and Thompson (1987). The tree is con-

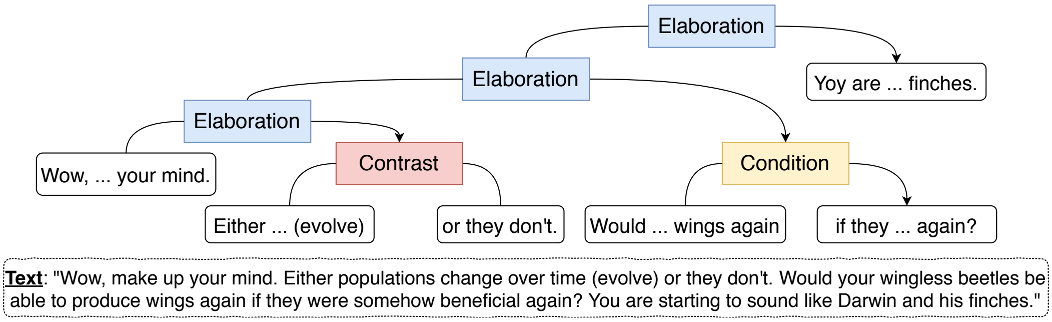

Elaboration
Elaboration
Yoy are ... finches.
Elaboration
Contrast
Condition
Wow, ... your mind.
Either ... (evolve)
or they don't.
Would ... wings again
if they ... again?
Text: "Wow, make up your mind. Either populations change over time (evolve) or they don't. Would your wingless beetles be able to produce wings again if they were somehow beneficial again? You are starting to sound like Darwin and his finches."

Figure 1: Discourse tree for text from an Internet forum.

structed step by step from the leaves to the root.

Initially, the text is divided into several intervals, called elementary discourse units (EDUs). Each of them contains a single thought, which cannot be broken more. Further, these intervals are connected by discourse relations such as "Elaboration", "Joint" and "Condition". After the unification of the elementary units, there are formed larger intervals of the text, which can be also connected by the corresponding discourse relations. This process can be continued until the only one node will remain (the root of the tree).

RST identifies two types of vertices: "Nucleus" and "Satellite". Vertices of the first type contain the crucial parts of the text, whereas, vertices of the second type provide some additional information.

Figure 1 demonstrates an example of a discourse tree for a text from an Internet forum.

3.2 EDU Embeddings

The pretrained Deep Averaging Network was chosen to construct embeddings of elementary discourse units (text spans). This model is a variation of the Universal Sentence Encoder, proposed by Cer et al. (2018). DAN averages word embeddings and applies a stack of fully-connected layers to get the final vector representation of the text.

We also consider parts-of-speech tags as additional information about the text. We embed POS-tags as vectors using one-hot encoding.

The final vector representation of an EDU is the concatenation of a semantic embedding from DAN and syntactic embedding constructed due to the POS-tags.

3.2.1 Recursive Neural Network

A recursive neural network encodes a tree as a vector of a fixed dimension. Similar to the tree construction in RST, the encoding occurs recursively along subtrees from leaves to root. The process of obtaining an embedding of a subtree with the root in the node i can be described as follows.

Let x_i denote the text embedding corresponding to the node i.

$$x_i = \begin{cases} \text{EDU embedding, if } i \text{ is the leaf} \\ \text{Embedding of the empty text, else} \end{cases}$$

Text Encoder applies a fully-connected layer to this pre-trained vector:

$$\text{Text\_Enc}(i) = \text{FC}(x_i) \tag{1}$$

Let nodes denoted as j and k be children indices for the node i, and r be the name of the discourse relation that characterizes the link between them. Dummy child vertices containing empty text are added for the leaves. The vector representation of the input associated with i concatenates four vectors as follows:

$$t_i = \text{Concat}[\mathbb{I}[j \text{ is Nucleus}], \mathbb{I}[k \text{ is Nucleus}], \text{OneHot}(r), \text{Text\_Enc}(i)] \tag{2}$$

In (2) $\mathbb{I}$ is the indicator function.

An embedding of the tree which has root in the node i is computed based on embeddings of its left and right subtrees due to the binary TreeLSTM model (Tai et al., 2015).

$$h_i = \text{TreeLSTM}(t_i, h_j, h_k) \tag{3}$$

We use TreeLSTM with dropout regularization of recurrent networks suggested by Semeniuta et al. (2016). Formally, the model is expressed with equations (4), (5) and (6).

$$\begin{pmatrix} \boldsymbol{i}_i \\ \boldsymbol{f}_{i0} \\ \boldsymbol{f}_{i1} \\ \boldsymbol{o}_i \\ \boldsymbol{u}_i \end{pmatrix} = \begin{pmatrix} \sigma(W_i[t_i, h_j, h_k] + b_i) \\ \sigma(W_{f_0}[t_i, h_j, h_k] + b_{f_0}) \\ \sigma(W_{f_1}[t_i, h_j, h_k] + b_{f_1}) \\ \sigma(W_o[t_i, h_j, h_k] + b_o) \\ D(\tanh(W_u[t_i, h_j, h_k] + b_u), \alpha) \end{pmatrix} \tag{4}$$

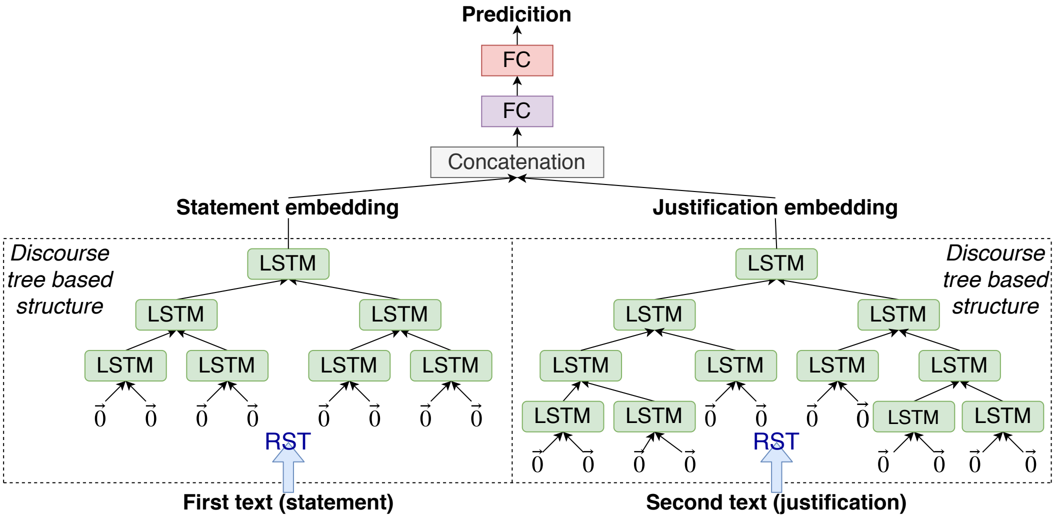


Figure 2: DSNDM model. Here, "LSTM" applies the TreeLSTM cell. Cells with the same color use the same weights. Each cell applied to EDUs receives zero vectors as embeddings of its children.

$$c_i = c_j * \boldsymbol{f}_{i0} + c_k * \boldsymbol{f}_{i1} + \boldsymbol{i}_i * \boldsymbol{u}_i \tag{5}$$

$$h_i = \boldsymbol{o}_i * c_i \tag{6}$$

Here, σ is the sigmoid function, D is the Dropout function, α is the dropout rate and $*$ is the element-wise multiplication. The memory cell is denoted as c. There are two forget outputs since the trees are binary.

The embedding received at the root of the tree is the vector representation of the entire text.

3.3 DSNDM

We propose DSNDM - siamese model based on the recursive neural network. There are two stages of the final model.

Firstly, the embeddings of the discourse trees for each of the input texts are calculated. The trainable parameters for both texts are the same. At the next stage, the resulting embeddings are aggregated for solving the categorization task. Here, the model concatenates the calculated trees embeddings and applies a sequence of two fully-connected layers to it. The last layer utilizes the Softmax function to map input features to the class probability space.

The main advantage of the proposed model is that it is capable of end-to-end learning. Figure 2 shows the architecture of the model. In this case, it solves the fact-checking problem. At the same time, it is almost the same for question-answer systems, except its inputs: the first text is a question, and the second is an answer.

3.4 Integration of the Attention Mechanism

We suggest a way of the integration of the attention mechanism (Vaswani et al., 2017), which has gained popularity in many NLP tasks. The main idea is that a constructed embedding of a question/statement can be used to filter information while constructing an embedding of an answer/justification. Thus, at each step, the model decides information from which subtree is more useful. The attention module can be integrated into the equations of the TreeLSTM model as follows.

Let us consider the Attention module, in which the key is the vector k and the values are represented by the matrix Q. In our case, the key is the embedding of the first text. The matrix Q is composed of vectors $q_1 = c_j * \boldsymbol{f}_{i0}$ and $q_2 = c_k * \boldsymbol{f}_{i1}$ and has the dimension $2 \times d$, where d is the dimension of the memory vector. Then, instead of equation (5), the memory cell vector is recalculated using attention matrices:

$$c_i = 2 \cdot \mathrm{Att}(k, Q) + \boldsymbol{i}_i * \boldsymbol{u}_i \tag{7}$$

$$\mathrm{Att}(k, Q) = \mathrm{SM}\left(\sum_{j=1}^{|Q|} \langle W_K k, W_Q, q_j \rangle W_V q_j\right) \tag{8}$$

Here, SM is the Softmax layer which is used for normalization. In (7), multiplication by 2 is necessary to maintain a balance with equation (5). In (8), matrices W_K, W_Q and W_V are trainable matrices of parameters of the Attention module.

Equation (7) is utilized instead of (5) only to construct the embedding of the second text.

3.5 Training Techniques for Ranking

DSNDM can be used both in the text classification task and in the ranking task. In this paper, we investigate three ranking techniques.

1) Classification-based

All pairs in the dataset can be divided into two groups based on relevance. The suggested model can be applied to solve the binary classification of text pairs with these groups. The ranking of the answers for each question is carried out using the class probabilities predicted by the model. The architecture of the model completely coincides with the base one in this case, and cross-entropy loss is used to train it.

2) Pointwise ranking

In this case, the main task is the regression problem. Let $\{(q_i, a_i)_{i=1..N}\}$ is the set of the given pairs, and $\{r_i\}$ are the corresponding relevance scores. Let the proposed model is denoted as DSNDM(q, a, w), where w are model parameters. Then, the model minimizes the following loss:

$$\sum_{i=1}^{N} (\text{DSNDM}(q_i, a_i, w) - r_i)^2 \to \min_{w} \tag{9}$$

3) Pairwise ranking

Here, the input are triplets $\{(q_i, a_i^+, a_i^-)_{i=1..M}\}$, where the relevant and irrelevant answer are selected for each question. These triplets can be generated from pairs using relevance scores. The ranking model solves the regression problem and minimizes the loss from (11).

$$\text{PN(w)} = \text{DSNDM}(q_i, a_i^+, w) - \text{DSNDM}(q_i, a_i^-, w) \tag{10}$$

$$\sum_{i=1}^{M} \frac{1}{1 + \exp(\text{PN(w)})} \to \min_{w} \tag{11}$$

4 Results

4.1 Automatic Fact Verification

4.1.1 LIAR-PLUS dataset

This dataset (Alhindi et al., 2018) contains the statements of politicians collected from politifact.com and labeled by experts, depending on the veracity on a 6-point scale. Binary classification is also possible when all labels less than four indicate lie and the rest indicate truth. The LIAR-PLUS dataset is an extension of the LIAR dataset. It contains automatically extracted justification for each statement.

The dataset also contains metadata with information about the politician and the global context of the statement. The LIAR-PLUS dataset can be used in four scenarios, depending on the restriction on the available data: S (only statement is used), S + M (statement and metadata), SJ (pairs: statement and justification), and S + JM (all available data). The model proposed in this paper is applied to pairs in the SJ scenario. At the same time, the model can be also used in the S scenario utilizing only the recursive neural network.

The dataset contains 12,782 statements which were split into the train, validation and test samples in the ratio of 10:1:1. This dataset is balanced, and the accuracy metric can be used to compare results.

4.1.2 Implementation Details

Firstly, text preprocessing was applied. We converted texts to lower case, removed extra characters and stop words. The open-source discourse parser ALT (Joty et al., 2012) was applied to the prepossessed texts to obtain discourse trees. Finally, the constructed trees were converted to the format described in section 3.1.

We used the DyNet python library to implement our model. The size of the hidden layer in LSTM cells was established at 100, the dropout rate α at 0.1, the learning rate at 0.004 and the number of units in the fully-connected layer in the Text Encoder at the dimension of x_i. We chose the Adagrad optimizer which is less prone to overfitting for the assigned task. The optimal number of epochs is 4-9. The model was trained by mini-batches of 150 pairs of texts.

4.1.3 Experiments

The parser identified 18 unique discourse relations. The most popular relations are “Elaboration” (is chosen by default), “Attribution”, “Joint” and “Same-Unit”. Usually, the trivial relations are popular in texts, and the ALT parser tends to use it in uncertain cases.

We investigated the difference between relation distributions for the instances in “true” and “pants-fire” classes. The “Joint” relation is less common for truthful statements than for misleading statements (relative frequencies are 0.064 and 0.073). Thus, politicians tend to construct longer, complex sentences in the case of the deceptive statements. Besides, the “Attribution” relation is used more

Model	Binary		Six-way	
	valid	test	valid	test
LR	0.68	0.67	0.37	0.37
SVM	0.65	0.66	0.34	0.34
BiLSTM	0.70	0.68	0.34	0.31
P-BiLSTM	0.69	0.67	0.36	0.35
DSNDM	**0.71**	0.69	**0.40**	0.40
DSNDM + Att.	0.70	**0.71**	**0.40**	**0.41**

Table 1: Model performance (macro-avg. F1-score) on the LIAR-PLUS dataset.

often for truthful statements (frequencies are 0.17 and 0.15). In the biggest part of cases, it indicates a link to the source. Thus, the relations contain some important information by themselves.

We compared the model with the methods proposed in (Alhindi et al., 2018). In addition to well-known baselines (such as linear regression and SVM), BiLSTM and P-BiLSTM are considered. The last one is the siamese model based on the BiLSTM architecture. Table 1 demonstrates the results for the 6-class and binary categorization tasks.

The table shows that the DSNDM model significantly improves the results of baselines, especially in the case of the multiclass classification.

The fully-connected layer in the Text Encoder is crucial since it adds up to 0.02 to accuracy. The usage of the POS-tags embeddings also improves the overall quality approximately by 0.003-0.01.

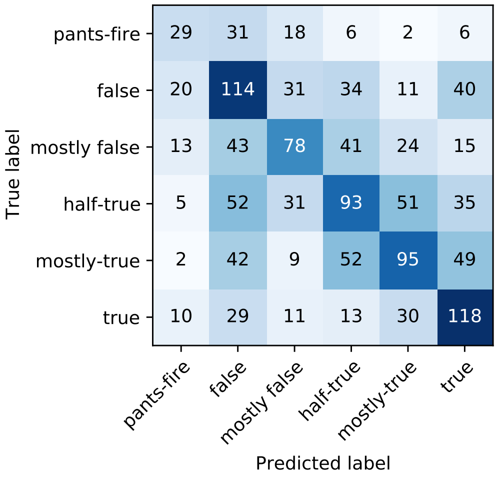
The DSNDM model with the integrated attention module (denoted as DSNDM + Att.) reached the best results for the test set. This improvement is not significant because of the binary structure of trees (the attention module re-weights only two vectors at each node).

4.1.4 Error Analysis

It is worth emphasizing that in some case trees for statements contain only one node. Therefore, discourse analysis does not suffice to categorize it. For the deepest trees which contain more than 45 nodes in the statement and justification in total (there are 89 such instances in the dataset), the F1-score metric is higher than 0.46.

The confusion matrix is shown in Figure 3. It demonstrates that DSNDM mainly intermingles close labels. However, at the same time, it confuses the classes "false" and "true" in some cases.

We distinguish several types of such instances which are demonstrated in Table 3 (see Appendix A). Firstly, there are some cases when refutation partly repeats the statement. Then, the model with attention focuses mainly on the repeated part and marks the misleading statement as "true". Secondly, the justification text can be extracted inaccurately and be not sufficient to estimate the veracity of the statement. Apart from that, the justification can be complex and contain only one useful sentence like in the third example. Finally, in the last pair, justification indicates that the statement can be labeled as "false" in some general cases, but has the label "true" in the considered case. Therefore, this justification contains useless thoughts and can be provided more accurately.



	pants-fire	false	mostly false	half-true	mostly-true	true
pants-fire	29	31	18	6	2	6
false	20	114	31	34	11	40
mostly false	13	43	78	41	24	15
half-true	5	52	31	93	51	35
mostly-true	2	42	9	52	95	49
true	10	29	11	13	30	118

Figure 3: Confusion matrix for DSNDM + Att. model for the LIAR-PLUS dataset.

Thus, the quality of the proposed model is limited by several factors: the size of the discourse trees, the quality of the discourse parser, and the quality of the provided justifications.

4.2 Question Answering Systems

4.2.1 ANTIQUE Dataset

This dataset (Hashemi et al., 2020) contains non-factoid questions with a set of possible answers for each of them. The authors selected questions from the Yahoo! Webscope L6 (nfL6) database. The questions were preliminary filtered: short questions, duplicates, and some complex cases were removed.

The corpus contains 2,426 questions in the training sample and 200 in the test sample. Answers for each question were selected both from the question forum thread and from other threads using the BM25 algorithm. In this way, 27,422 answers were allocated for training, and 6,589 instances for testing.

The resulting QA pairs were labeled on a 4-point scale depending on the relevance of answers using the crowdsourcing procedure. The authors also proposed a binary classification task where instances with labels 1 and 2 can be considered as irrelevant, and instances with labels 3 and 4 can be considered as as relevant. Thus, the most common ranking metrics such as MAP and MRR can be used in the second task. At the same time, the multiclass metric nDCG can be also considered. The number of the best answers for questions differs, but on average it is approximately equal to 8.

The dataset is not balanced: the number of relevant answers is almost twice bigger than irrelevant ones. The authors used a negative sampling procedure to train baseline models, increasing the size of the dataset several times. However, it is important to emphasize that these additional QA pairs were not included in the publicly available dataset.

Questions are not very long and contain about 11 words on average. At the same time, the answers are much longer and contain more than 47 words on average. Therefore, it can be problematic to use the standard BERT model, but it is an advantage for the discourse analysis.

4.2.2 Implementation Details

The implementation details are almost the same as described in Sect. 4.1.2 except for some hyperparameters. It is better to choose the smaller dimension of the hidden vectors. The dimension of vectors in TreeLSTM was set to 100, and in the TextEncoder layer was set to 64. It takes 1-3 epochs to achieve optimal quality. A tenth of the training set was used as a validation sample during training.

4.2.3 Experiments

The discourse parser identified 18 different discourse relations like in the first task. However, in this case, the frequency statistics of relations are very similar for different classes. It is due to the fact that in this task the second text (answer) is not auxiliary.

We compared the suggested model with the baselines presented in (Hashemi et al., 2020). It should be highlighted that these baselines were trained on the extended dataset. The authors additionally performed the negative sampling procedure. Therefore, it is not correct to compare the results obtained on the available base dataset with the results obtained on the extended dataset.

	Model	**MRR**	**P@1**
1	BM25	0.4885	0.3333
	DRMM-TKS (2016)	0.5774	0.4337
	aNMM (2016)	0.6250	0.4847
	BERT (2018)	***0.7968***	***0.7092***
2	ConvKNRM [pairwise]	0.4920	0.3650
	BERT [pointwise]	0.6694	0.5550
	BERT [pairwise]	***0.6999***	***0.5850***
	Tuned BM25	0.5802	0.4550
	Tuned SDM	0.5377	0.4400
3	Base [classif.]	0.6792	0.5350
	+ Att. [classif.]	0.6830	0.5350
	Base [pointwise]	0.6864	0.5300
	+ Att. [pointwise]	0.7098	0.5650
	Base [pairwise]	0.7120	0.5800
	+ Att. [pairwise]	**0.7267**	**0.6000**

Table 2: Model performance on the ANTIQUE test set. 1: Models presented in (Hashemi et al., 2020), 2: Models presented in (MacAvaney et al., 2020), 3: DSNDM

Apart from that, we considered several models discussed in (MacAvaney et al., 2020). In this paper, several negative examples were also added for each question. However, they were most likely selected only from the training corpus, since the authors were unable to reproduce the BERT results from the original paper.

MacAvaney et al. (2020) proposed various modifications of the training loss by adding a weight for each pair. We do not compare with the results obtained with a modified curriculum since we consider only the basic pointwise and pairwise losses.

The comparison results are presented in Table 2. It shows that DSNDM + Att. model trained using pairwise loss achieves high MRR and P@1 metrics. Its results are superior to the results of the best BERT model presented in (MacAvaney et al., 2020). We also trained BERT ourselves and obtained results close to it, and we could not reproduce the results from the original paper too. Also, the pointwise ranking performed better than the classification-based method.

The attention mechanism improved quality in all cases, especially for the pointwise and pairwise techniques.

4.2.4 Error Analysis

We investigated the mistakes of DSNDM trained for the classification problem. Figure 4 shows the

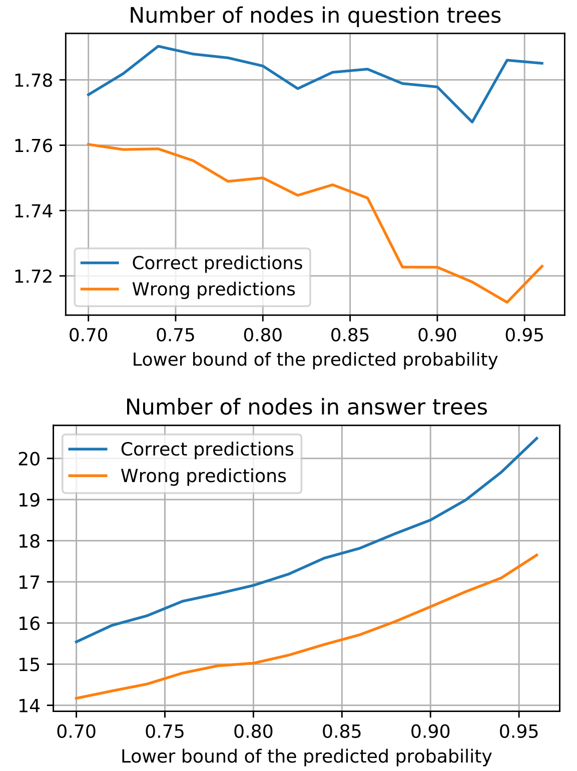


Figure 4: Dependence of the number of nodes in the discourse trees of questions (on the top) and answers (on the bottom) on the confidence of DSNDM in cases of correct and wrong predictions.

dependence of the average number of nodes in questions/answers on the confidence of the model. Thresholds are moved along the horizontal axis. Statistics are calculated only for pairs for which the model predicts a probability that exceeds the selected threshold. One can see that for both questions and answers, the number of nodes in discourse trees for correctly classified pairs is greater than for incorrectly classified ones. Thus, DSNDM makes wrong predictions mostly for small trees. Also, the plot for questions demonstrates that the model's greater confidence in the wrong answer is frequently triggered by the smaller size of the question tree. Therefore, the quality of the proposed model is closely related to the size of the discourse trees for this task too.

In this case, we distinguish several typical mistakes which are demonstrated in Table 4 (see Appendix A). In the first pair, the question contains only a few significant keywords, and the model focuses mainly on them. Despite the fact that the answer is irrelevant and unrelated to the question area, it often uses the same keywords. Thus, similar EDU embeddings do not contribute to the correct classification. In the second example, the meaning of the answer and the question is the opposite. That is, despite the correctness of the answer, its text refutes the information in the question. If the question contains only one node, then such instance is one of the most difficult for analysis. Finally, the last example demonstrates that in some cases the correct answers may be formulated in the way not expected by the authors of the questions. Thus, the quality of the model is also limited by the variability of possible answers.

5 Conclusion and Future Work

In this paper, we investigated the utility and importance of the discourse analysis for text pairs categorization and ranking. We considered two typical tasks in which discourse analysis seems promising: automatic verification of political statements and ranking in question answering systems.

We modified TreeLSTM to effectively encode discourse trees and proposed DSNDM which is capable of processing pairs of texts. In addition, the integration of the attention mechanism in the proposed model was suggested to obtain more useful embeddings of subtrees. Moreover, we investigated three training techniques for the ranking task.

The experiments were performed on the LIAR-PLUS and ANTIQUE datasets. DSNDM efficiently learned the match between discourse tree structures and achieved high quality in both tasks. Besides, the attention module improved the metrics of the base model in all cases. The error analysis showed that the model processes deeper trees more successfully.

There are possible directions for future work: the use of trees not only of a binary structure, the modification of vector representations of EDUs, as well as the investigation of the performance of DSNDM in other various tasks where discourse analysis may be helpful, e.g. machine translation, chat-bots and other QA systems. Apart from that, we will experiment with other hierarchical structures (e.g. syntactic) for deeper analysis of the importance of the RST-based structure in the proposed model.

Acknowledgments

The article was prepared within the framework of the HSE University Basic Research Program and funded by the Russian Academic Excellence Project '5-100'.

References

Tariq Alhindi, Savvas Petridis, and Smaranda Muresan. 2018. Where is your evidence: Improving fact-checking by justification modeling. pages 85–90.

Aimée Alonso-Reina, Robiert Sepúlveda-Torres, Estela Saquete, and Manuel Palomar. 2019. Team GPLSI. approach for automated fact checking. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 110–114, Hong Kong, China. Association for Computational Linguistics.

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from RST discourse parsing. *CoRR*, abs/1509.01599.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder. *CoRR*, abs/1803.11175.

Giovanni Luca Ciampaglia, Prashant Shiralkar, Luis Mateus Rocha, Johan Bollen, Filippo Menczer, and Alessandro Flammini. 2015. Computational fact checking from knowledge networks. *PLoS ONE*, 10.

Ryan Clancy, Ihab F. Ilyas, and Jimmy Lin. 2019. Scalable knowledge graph construction from text collections. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 39–46, Hong Kong, China. Association for Computational Linguistics.

Wanyun Cui, Yanghua Xiao, Haixun Wang, Yangqiu Song, Seung-won Hwang, and Wei Wang. 2017. Kbqa: Learning question answering over qa corpora and knowledge bases. *Proceedings of the VLDB Endowment*, 10:565–576.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Boris Galitsky. 2019. *Learning Discourse-Level Structures for Question Answering*, pages 177–219.

Boris Galitsky and Dmitry Ilvovsky. 2016. Discovering disinformation: discourse level approach. In *15th National Conference on Artificial Intelligence with International Participation (CAI)*, pages 23–32.

Boris Galitsky, Dmitry Ilvovsky, and Sergei O. Kuznetsov. 2015. Text classification into abstract classes based on discourse structure. In *RANLP*, pages 200–207.

Jiafeng Guo, Yixing Fan, Qingyao Ai, and W. Croft. 2016. A deep relevance matching model for ad-hoc retrieval. pages 55–64.

Helia Hashemi, Mohammad Aliannejadi, Hamed Zamani, and W. Bruce Croft. 2020. Antique: A non-factoid question answering benchmark. In *Advances in Information Retrieval*, pages 166–173, Cham. Springer International Publishing.

Yangfeng Ji and Noah A. Smith. 2017. Neural discourse structure for text categorization. *CoRR*, abs/1702.01829.

Shafiq Joty, Giuseppe Carenini, and Raymond Ng. 2012. A novel discriminative framework for sentence-level discourse analysis. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 904–915, Jeju Island, Korea. Association for Computational Linguistics.

Aiting Liu, Ziqi Huang, Hengtong Lu, Xiaojie Wang, and Caixia Yuan. 2019. *BB-KBQA: BERT-Based Knowledge Base Question Answering*, pages 81–92.

Sean MacAvaney, Franco Maria Nardini, Raffaele Perego, Nicola Tonellotto, Nazli Goharian, and Ophir Frieder. 2020. Training curricula for open domain answer re-ranking. pages 529–538.

William Mann and Sandra Thompson. 1987. Rhetorical structure theory: A theory of text organization.

Yixin Nie, Haonan Chen, and Mohit Bansal. 2019. Combining fact extraction and verification with neural semantic matching networks. In *AAAI*.

Stanislau Semeniuta, Aliaksei Severyn, and Erhardt Barth. 2016. Recurrent dropout without memory loss. *CoRR*, abs/1603.05118.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. *CoRR*, abs/1503.00075.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The fact extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, pages 1–9, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need.

You Wu, Pankaj K. Agarwal, Chengkai Li, Jun Yang, and Cong Yu. 2017. Computational fact checking through query perturbations. *ACM Transactions on Database Systems (TODS)*, 42:1 – 41.

Liu Yang, Qingyao Ai, Jiafeng Guo, and W. Croft. 2016. anmm: Ranking short answer texts with attention-based neural matching model. pages 287–296.

A Appendix

Appendix contains typical examples of pairs for which DSNDM got wrong predictions.

Statement	Justification	Label
In rural Virginia, Sen. Warner ran 8 - 10 points ahead of a traditional Democrat – ahead of Senator Kaine, ahead of Governor McAuliffe.	Hallock said Warner, in this fall's Senate election, ran 8 - 10 points ahead of past performances by fellow Democrats McAuliffe and Kaine in rural Virginia. McAuliffe's portion of the rural vote, in his 2013 gubernatorial victory, was 3.6 percentage points below Warner's. Kaine's slice of the rural vote, in his 2012 Senate win, was 2.4 percentage points above Warner's.	false
In the U. S. Constitution, theres a little section in there that talks about life, liberty and the pursuit of happiness.	No court makes a legal decision based on the Declaration of Independence, Wilkes said. With his first speech as a bona fide candidate, Cain joins a long, bipartisan line of presidential hopefuls who have succumbed to foot - in - mouth disease. They include Cain's foe, President Barack Obama, who accidentally said there were 57 states during the 2008 campaign and U. S. Sen. John McCain, who said in an interview he was unsure how many houses he owned. Welcome to the 2012 presidential election season, folks.	false
In the Illinois Legislature, Barack Obama voted present, instead of yes or no on seven votes involving abortion rights.	Two other large groups, NARAL Prochoice America and Planned Parenthood, are not endorsing. Planned Parenthood, however, has given both candidates 100 percent ratings for their records on abortion. We stipulate that there are clearly different interpretations of the significance of Obama s present votes. But there s no doubt he made them.	true
As a result of climate change, ice fishermen in Wisconsin are already noticing fewer days they can be out on our ice covered lakes.	Not what is going to happen this year. Our rating It's been a longer and colder winter than in recent years. But that doesn't erase a trend that's been well - established. The number of days that the lakes have ice on them – making them safe for ice fishing – has declined.	true

Table 3: Typical mistakes of DSNDM on the LIAR test set where the model confuses "true" and "false" instances.

Question	Answer	Label
how does disneyland make it snow?	Well if you are using snow, just lay on you back in it and move your arms from your sides to the top of you head and open and close you legs a few times... to make snow angels!!!!	Out of context
Why cant teenagers vote?	Teens can vote. When theyre 18 and 19... Teens still have ALOT to learn. There is nothing wrong or demeaning about this. Even most ADULTS have alot to learn about politics. They go into a voting booth having no idea what party stands for what, or what candidate believes in what, and vote Democrat when their beliefs are Republican, or vote Republican when their beliefs are Democrat.	Correct answer
Why is Gordon Ramsey so popular?	Is he that popular? what little I have seen of him every second word is a swear word, if that makes him popular then it says a lot about what is wrong in this country... the man is a cretin.	Correct answer

Table 4: Typical mistakes of DSNDM on the ANTIQUE test set where the model confuses "Correct answer" and "Out of context" instances. .